Exploring Business Frequency Near Public Bike Transportation Stations

Makana Burch, ’25, Data Science Major Capstone

Question

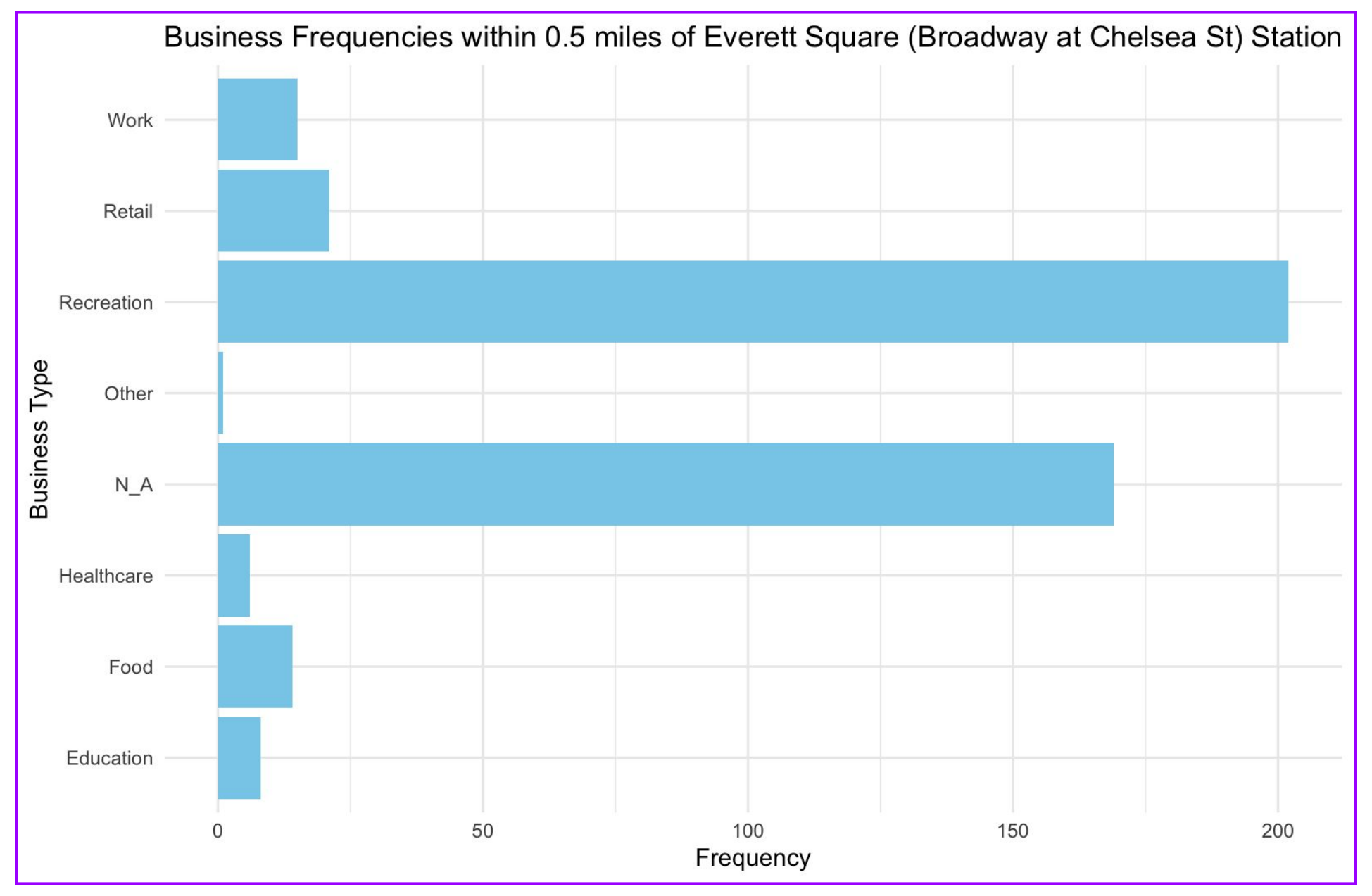
Is there a relationship between the proportion of subscriber rides to total rides ending at a station and the types of businesses within a half-mile radius of that station (office, food, recreation, etc.) in 2019?

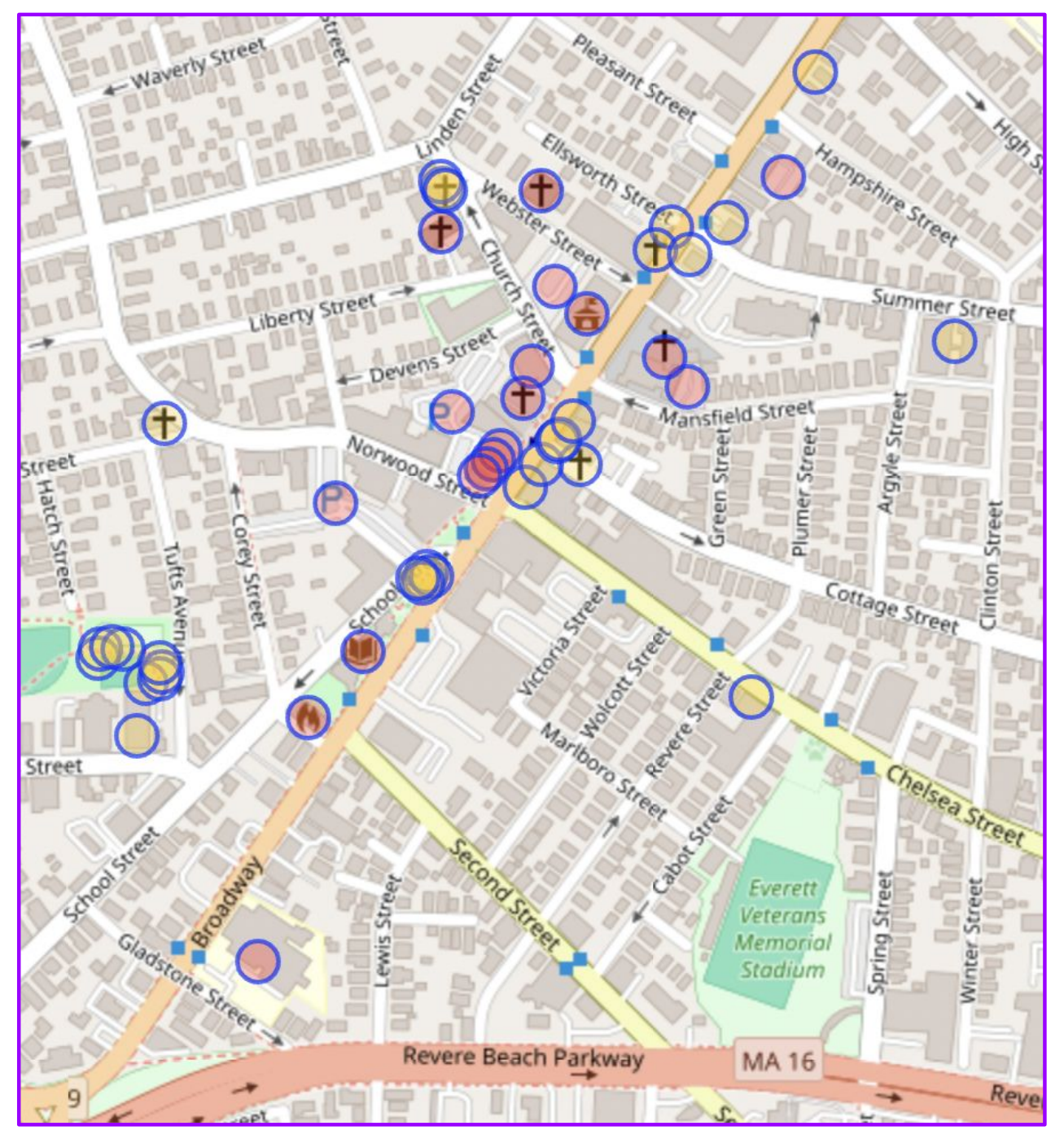
Introduction

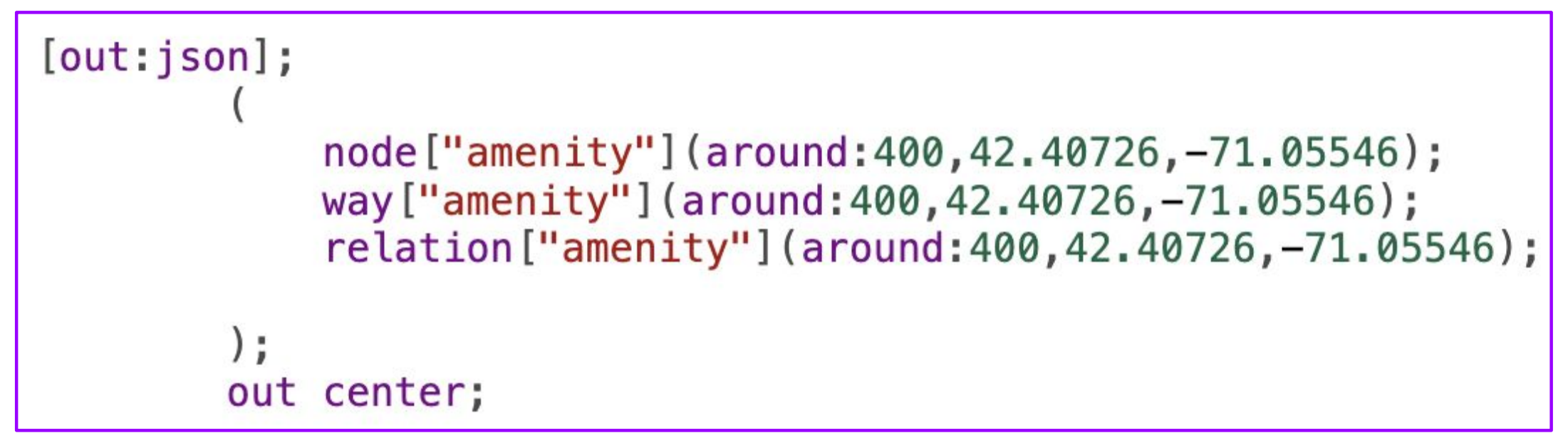
- Research conducted by Ghorbanzadeh, et. al. indicates that in choosing locations of subway stations, the most influential criterion are the population density around the potential station, its proximity to medical centers, and the soil type of the area.
- Data sources:
  - BLUEbikes is a public bike share service within the greater Boston area
    - Users can be subscribers or casual riders
      - Subscribers pay monthly or yearly with unlimited trips
      - Casual riders pay by the trip
  - Overpass Turbo is an online query service that uses OpenStreetMaps data to search the area around a given latitude and longitude for buildings and features that match the given query parameters
- Evidence of a relationship could
  - Allow BLUEbikes to consider business frequencies when considering new station locations
  - Could expand upon Ghorbanzadeh, et. al.’s research to show multiple business types affect station use/location for different types of public transportation
  - Provide more information about how people are using BLUEbikes, which could help for marketing, updating existing stations, determining available plans, etc
- Subscriber proportions are calculated by dividing the number of trips that ended at a given station that were taken by subscribers with the total number of trips that ended at that station
- Business types: Food, Retail, Education, Recreation, Healthcare, Work
  - Other: businesses that don’t fall into the above categories
  - NA: businesses that don’t have a recorded category

Methods

1. Download 2019 trip data from BLUEbikes
2. Use Python to get unique stations and calculate the subscriber proportion
   a. Remove repair, maintenance, and warehouse locations recorded in data
3. Use R to query Overpass Turbo to get business data within 0.5 miles of a station
   a. 0.5 miles chosen as reasonable walking distance from a chosen station
   b. Stations with overlapping collection radii both count those businesses
4. Use R to create scatterplots comparing business type frequency to subscriber proportion
5. Calculate the percentage of each business type around a station
6. Transformations:
   a. Logit transformation on subscriber proportion
   b. Log transformation on raw business frequencies
7. Use R to run single/multiple linear regression
   a. Single linear regression for each business type
   b. Multiple linear regression with all variables
   c. Multiple linear regression with most significant variables
8. Compare model efficacies and draw conclusions from results

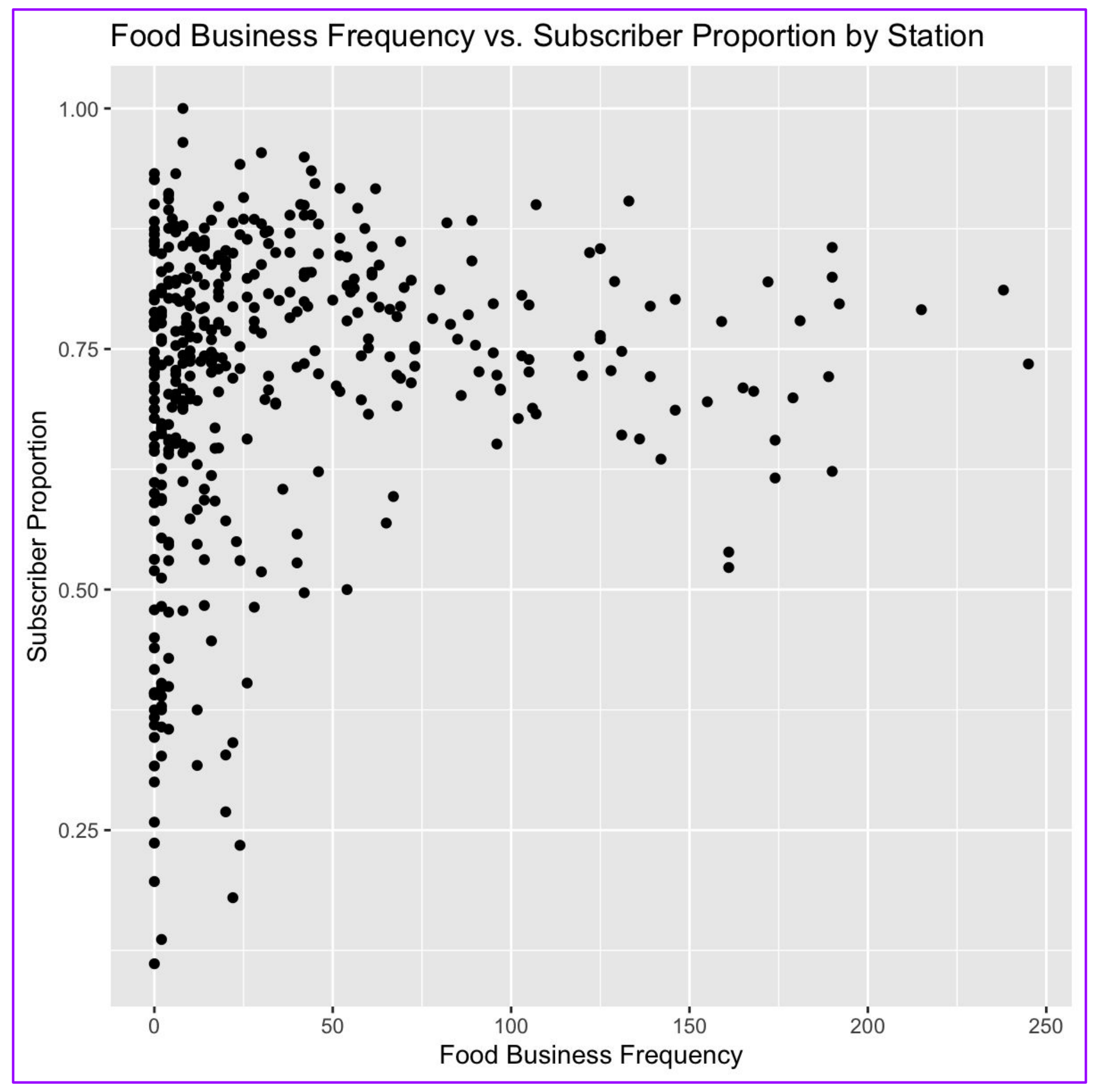




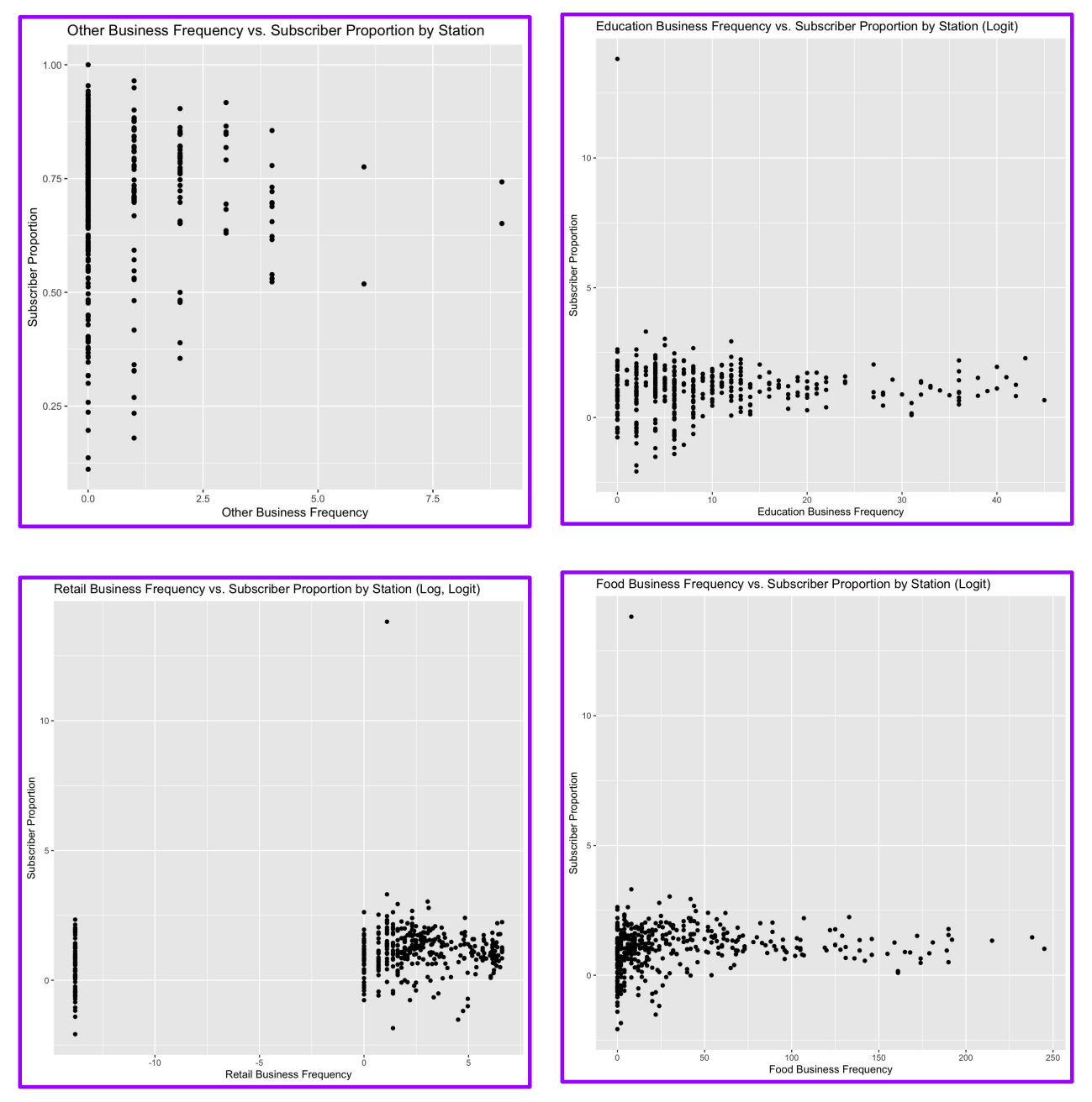


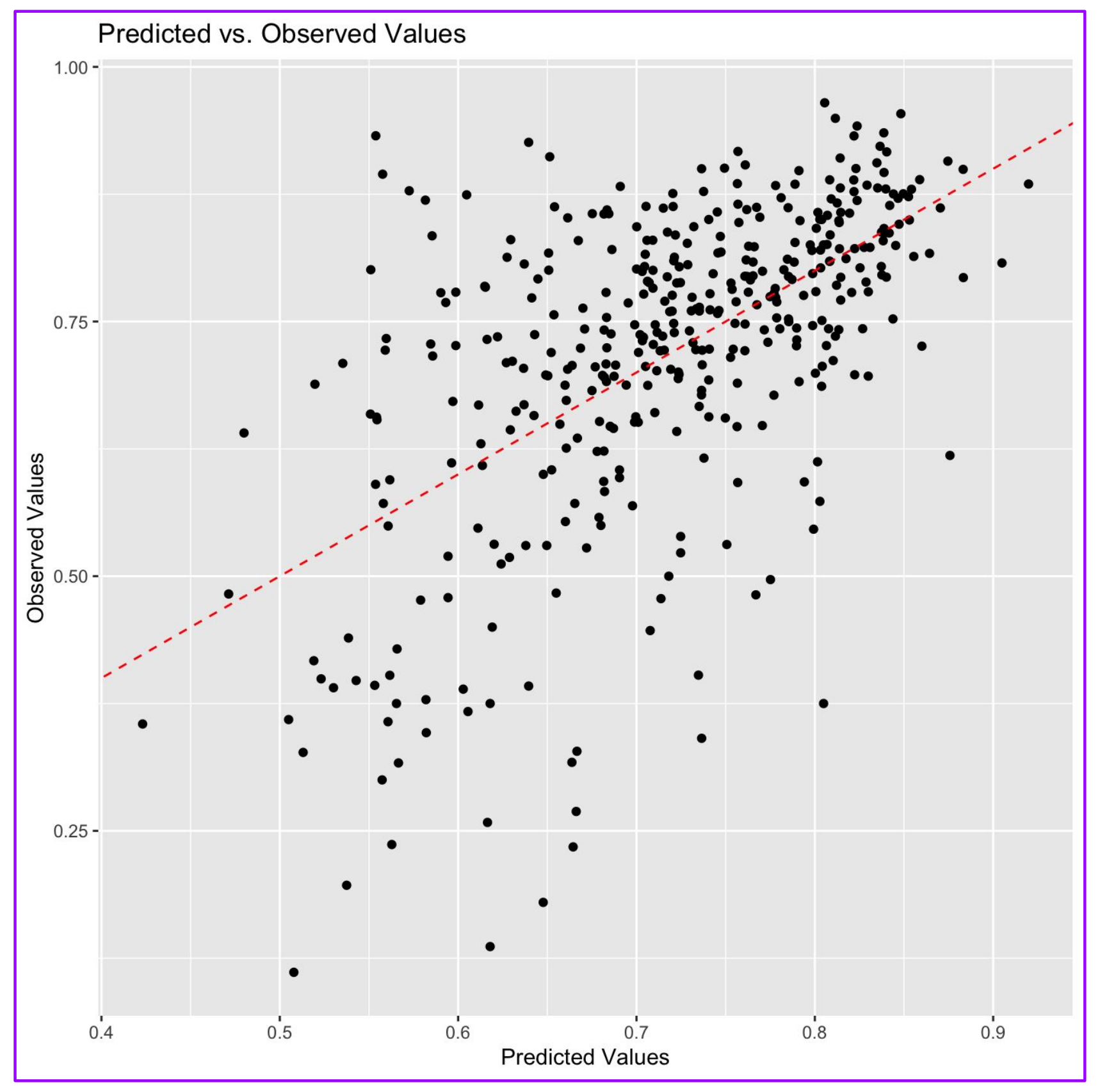
```
[out:json];
        (
            node["amenity"](around:400,42.40726,-71.05546);
            way["amenity"](around:400,42.40726,-71.05546);
            relation["amenity"](around:400,42.40726,-71.05546);

        );
        out center;
```

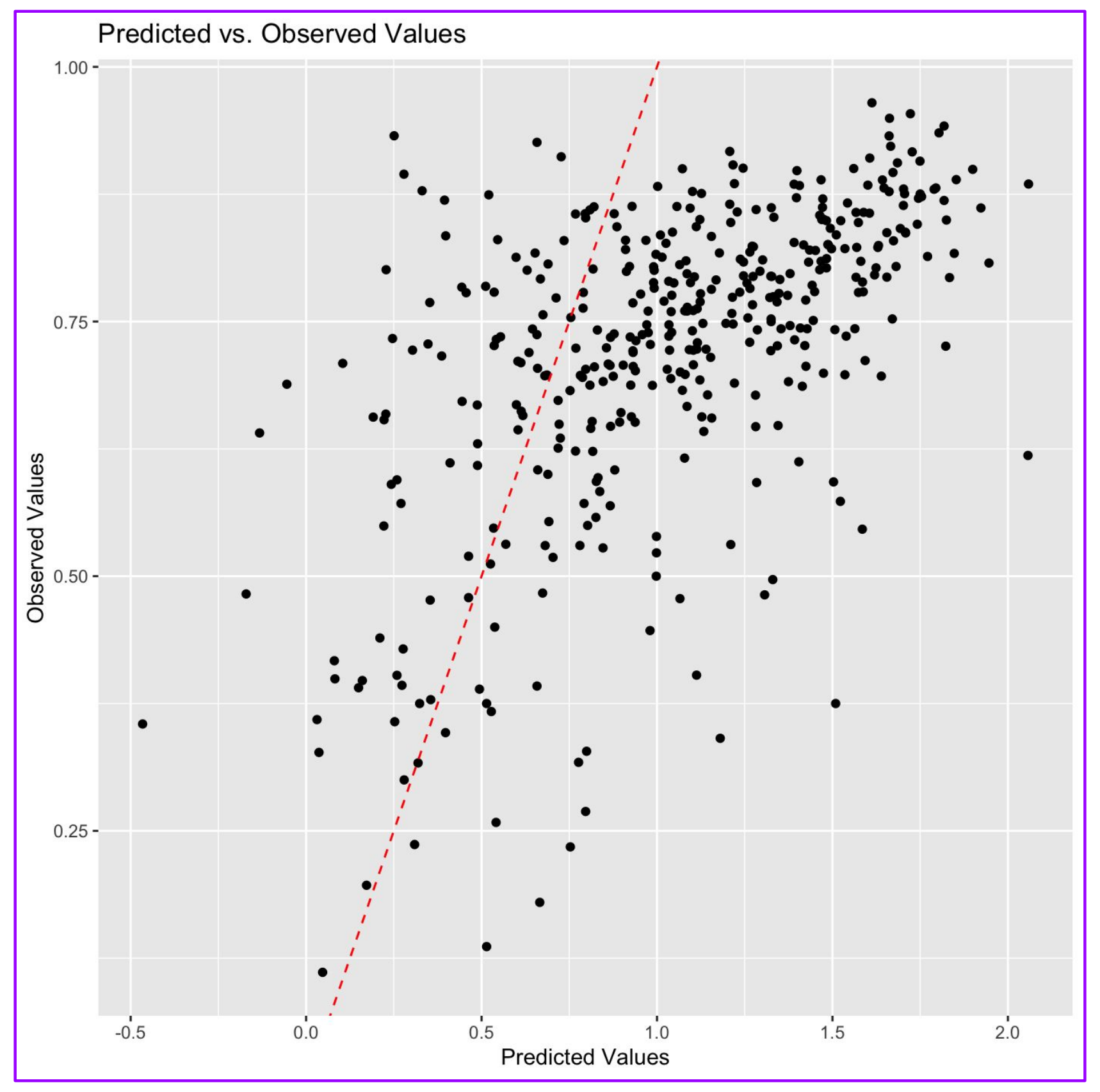
Bar Chart: Raw frequencies of business types
Text: Example Overpass Turbo query for Everett Square (Broadway at Chelsea St) Station
Map: Map output for query showing buildings that match the search criteria



Left: Scatterplot of raw food business frequency vs. subscriber proportion for every station
Below: additional example scatterplots







Left: Predicted vs. Observed values for all variable model predicting Subscriber Proportion

Right: Predicted vs. Observed values for all variable model predicting Logit(Subscriber Proportion)

Results

- Single Variable Models
  - Model with the highest adjusted R^2 (0.1283): Proportion~Log(Work)
  - Most models were statistically significant
  - Best individual models used: Log(Work), Log(NA), Log(Retail), Log(Food), Percent Recreation, and Log(Healthcare)
    - Determined based on a combination of adjusted R^2, BIC, and RMSE
- All Variable Models
  - Using Subscriber Proportion as dependent variable:
    - Significant variables: NA, Education, Retail, Log(NA), Log(Work), Percent Healthcare, Percent NA, Percent Work, Percent Education, and Percent Recreation
    - Adjusted R^2: 0.2924; P-value: <2.2e-16; RMSE: 0.128; BIC: -317.0433
  - Using Logit(Subscriber Proportion) as dependent variable:
    - Significant variables: Healthcare, NA, Education, Retail, Log(Work), Percent Healthcare, Percent NA, Percent Work, Percent Education
    - Adjusted R^2: 0.2918; P-value: <2.2e-16; RMSE: 0.654; BIC: 961.6871
- Multivariable Models
  - Subscriber Proportion:
    - Adjusted R^2: 0.2654; P-value: <2.2e-16; RMSE: 0.132; BIC: -420.3171
  - Logit(Subscriber Proportion):
    - Adjusted R^2: 0.2043; P-value: <2.2e-16; RMSE: 0.706; BIC: 939.6555
- Assumptions
  - Linearity of Errors assumption may not be met due to time aspect of data
  - Constant Variance and Normality checked for all variable models - no violations

Conclusions

- Healthcare was a significant predictor in all the models, corroborating Ghorbanzadeh, et. al’s findings
- The multivariable models did not do significantly better than the models with all variables
- The all variable models shared most of their significant variables
  - Subscriber Proportion differed in Log(NA) and Percent Recreation
  - Logit(Subscriber Proportion) differend in Healthcare
- The all variable models did not have many of the best individual model variables as significant (highlighted in blue above)
- Overall, business frequency information could be used to describe around 25% of the variation in subscriber proportions
- There is some relationship between the proportion of subscriber rides to total rides ending at a station and the types of businesses within a half-mile radius of that station
  - However, other factors should be considered when making decisions about new station locations or marketing/upgrading existing stations

References

Ghorbanzadeh, M., Effati, M., Gilanifar, M., Ozguven, EE., June 2020, Subway Station Site Selection Using GIS-Based Multi-Criteria Decision-Making: A Case Study in a Developing Country (article), Computational Research Progress in Applied Science & Engineering, Volume 06, Issue 02, pages: 60-69.

BLUEbikes, 2019, *Blue Bikes Comprehensive Trip Histories*, [csv], BLUEbikes, https://bluebikes.com/system-data.

WELLESLEY